A RETRIEVAL SYSTEM OF GREEK LEGAL DOCUMENTS

Angeliki Plati, Theodore Kalamboukis

Information Processing Laboratory,
Department of Informatics,
Athens University of Economics and Business,
76 Patission Street,
Athens 10434, Greece
ag.plati@gmail.com, tzk@aueb.gr

Abstract

An information retrieval system has been implemented, that serves the need for convenient, simple and fast information concerning legal documents. The retrieval of legal documents usually requires that the person concerned, should have specialized knowledge on the law and on the structure of the statutes, regulations, judicial decisions etc. Not every person has this knowledge on the law matters. The proposed system allows the user to search for legal information by asking simpler, non-standardized and free formed questions-queries.

The retrieved legal documents have the same form as the original ones and the system gives the option of navigating through the documents' pages. These facts may make this information retrieval system useful even to a lawyer.

1 Introduction

Information of text and its accessibility play a crucial role in the legal domain. The amount of available legal documents and other legal material, however, is enormous and continuously growing, making it more and more difficult to efficiently deal with it. Authors often speak about information overload [Wahlgren,1999] and the information crisis of law [Simitis,1970]. Thus, it is not surprising that the legal domain was one of the first fields where IR techniques were employed. Recently legal retrieval tracks have been introduced in the context of international conferences such as TREC [trec] and CLEF [clef].
The most important applications of IR techniques to the legal domain are
a) litigation support where IR techniques are used in searching large and heterogeneous data-sets. These data-sets may consist of e-mails, bills, reports and other legal material that might be important for a trial, and
b) computer-assisted legal research (CALR), where IR techniques are implemented in order to allow legal professionals to access legal sources via the computer.

There are some considerable retrieval systems of Greek legal documents that belong in the second category of the legal research. These retrieval systems are implemented in Greece and they are really useful to lawyers and professionals. Their applications specialize in law categorization (Civil Law, Criminal Law, Price Control Code etc). The user of these systems should know from the beginning in which “law-section” to look in before he starts his search.

The retrieval system of Greek legal documents, which is implemented within this project, belongs

also in the second category of the legal research but it could be mainly used by common people that they are not lawyers, legal practitioners etc. The legal documents that the system retrieves are in the form of "Official Government Gazette of the Hellenic Republic" (O.G.G.) [O.G.G.]

2 State of the art

2.1 National

There are some noteworthy retrieval systems of Greek legal documents. These platforms have been developed in Greece and they are mostly the legal professionals' "right hand". The aforesaid are "Nomos"[Nomos], "Legal Information Bank, Isokratis" of "Athens Bar Association"[Isokratis], "Digital Legal Library" [Digital Legal Library] etc. The accessibility to the most of these systems is not free, on the contrary they are quite expensive. In addition, "Nomos" is subject to time charge and "Isokratis" is free only for lawyers. Despite of their charges, they are really helpful to those who use them because of their great utilization of law categorization. Their applications provide complete thematic indexes of the legal material.

The above platforms are completely updated. When the user is searching for a specific legislation, the system brings back its latest modification, but there are also links at the top of the result pages that lead to the initial legislation. The searching brings back articles of the laws and it should be started by selecting a topic from the thematic index. The user can also search for articles by selecting the year, the issue and the number of the O.G.G. he is interested in. In addition, the user can start his research by choosing the European Law field. The above platforms give also the ability of textual searching but in any case, the user should be able to start the searching process by following one or more of the aforementioned ways. Otherwise, the textual searching may not be sufficient and the results may be redundant and indiscriminate.

It is obvious that if someone wants to use the applications that the aforementioned platforms provide, he should be aware of the law matters and the legal material. These systems are mostly useful to lawyers who encounter statutes on a daily basis. Additionally, many times the user-lawyer may need to combine two or more of these legal systems to find what he exactly needs because sometimes there is a lack of Government Circulars in some of them (e.g. "Nomos"). Moreover, "Nomos" for example, which is mainly used by the professionals, does not bring back results that contain internal links to the pages of the O.G.G. to which they belong. So, the user cannot navigate through an O.G.G.'s pages. This ability could be quite useful though. Furthermore, in "Nomos" textual searching may be complicated, occasionally. This happens because the user has to clarify to the system whether the words he is using in his research are related -or not- to each other, and this cannot be always clear. In addition, in "Nomos" once more, if the user desires to find not only the words written as such but also their derivatives or other words that include them, in the legal documents, then he has to write the words that he looks for by replacing the last letter of them with "*". So, this procedure may not be that manageable.

We can easily understand that a user, who is not a legal professional or a lawyer, would use a simple and effective textual searching application much more effortlessly than by initially choosing, for example, the kind of legislation or the kind of the code in order to move to the right direction and use a complicated textual searching, the way the above systems require. Usually, it is quite difficult for a user, who is not a legal practitioner and who is not acquainted to the legal matters, to comprehend which "part" of the law what he looks for belongs in. Thus, a simpler mechanism could be more effective. Moreover, links inside the legal documents indicating previous and next pages of the O.G.G.s would be very useful to the common user. The aforementioned platforms provide links indicating only relative legislation. The common user is likely to prefer simpler, more accessible, google-like and undoubtedly free or at least less expensive applications.

3 Pilot application

This application is a search engine for Greek legal documents, accessible and easy to use [Plati, 2012]. It is Google – like and it offers a simple and pleasant environment. The returned results of the search are pages of the respective O.G.G.s, in which probably the legal information, the user looks for, is found. Those pages have the same form like the ones in the original pdf document. This system provides the ability of searching legal documents by choosing – if required by the user – the number of the O.G.G. issue or/and the year of the publication and , surely, by performing textual searching with one or more key words.

The returned results of the search, which are in html form, satisfy the terms that are given, by choosing the issue and the year of the desirable O.G.G. and secondly they constitute the pages of the O.G.G.s that contain the most numerous occurrences of key words the user has written in the field of textual searching. The words written by the user don't need to be in a special form or contain special characters or symbols. The more often the term-occurrences are in one page, the "higher" this page is found in the list. Those words and their derivatives, as well as words including them, are highlighted inside the html pages of the results. Additionally, html pages contain links to the next or previous respective page (if there is any) in the O.G.G. where they belong, so that the user can easily navigate in it. There is, also, a link leading directly to the page of the total search results. Moreover, the page of the returned results provides a pager for navigating through the list of the results. There is, also, one more possibility, though. The user can, if desired, download the equivalent pdf document of the O.G.G. by clicking on the link located next to the one of the individual O.G.G. page, on the list of the results.

This system is a first step to the creation of a full, ergonomic and effective tool for people with little knowledge on legal matters and who need to find exactly what they ask for, simply by typing it. If this easy-to-use information retrieval system improves and is accomplished, it could satisfy the needs of the professionals and be absolutely complete and trustworthy.

In order to make the system functioning, a database was created, in which the O.G.G. file paths (both in html form and pdf form) have been stored. Their identities, their name and some other necessary information for the navigation and data are stored as well. There has also been an edit on all the pdf files by special tools in order to convert them in html pages, after being firstly converted to txt files, so that some important and necessary information is collected. Open source tools like these, are pdftohtml [pdftohtml] and HtmlAsText [HtmlAsText] respectively. Tailored to our problem software was also developed, as for example, the removal of hyphenation at the end of lines. There had been though several problems during the conversion of the files into various forms. This occurred due to the unstructured and not rigorous form of the original files. The original forms contain many scanned pictures, signatures, different text formats, for instance tables etc. Some other times, a steady structure and formatting are not maintained in every document. This, as it can be easily understood, toughens up almost all the conversions and consequently, later on, the creation of the indexes from the txt files, that followed. After these quite difficult preprocessing procedures, the creation of the web application followed, as well as the implementation of the textual searching in through the pages of a document and the procedure of the development of the graphic environment. Precious functions of the Lucene [Lucene] search engine have been used for the tokenization and the creation of indexes, in which the search is done, as well as for the search of key words that the user is looking for in them. Before the indexing, stopwords removal was applied and stemming, a procedure to reduce words to their morphological roots, by stripping off suffixes, with the help of GreekStemmer [Kalamboukis, 1995], concerning that the Greek language has many peculiarities. All the above mentioned tools, functions and libraries were optimized through the use of the Java [Java] programming language.

The retrieval was based on the classical vector space model [Manning, 2008] and the TF*IDF (Term Frequency times Inverse Document Frequency) weighting scheme was used for terms in the documents. To face the problem with the size variability of legal documents, which varies from one to several tens of hundreds of pages, we have applied retrieval on portions (passages) of the original texts. In this work we have define the size of a passage equal to a page. Passage retrieval [Rosso, 2011] apart of a kind of document normalization acts as a filter of non-relevant information because it reduces the original document collection to a set of passages in which the user information needs are satisfied. For a query, Q, and a document D, of k pages, $D=(P_1,P_2,\ldots,P_k)$, the score of relevance of the document is defined by:

$$SCORE(Q,D) = \sum_{P_j \in D} score(Q,P_j)$$

A page not retrieved by the query Q, will have zero score. The default similarity measure of the Lucene search engine was used to calculate to score of a passage, based on the cosine measure.

Concerning the functionality of the system, the user initially can choose, if desired the year or/and the issue of the O.G.G. considered to be related to the information required. Then, the only thing that the user can do is to type the key words that he thinks that are related to the topic of the legislation required and click on the button “Find”. It should be mentioned at this point, that the user could use the choices that are provided, individually. That is, he could if he would like to, to complete only the field of the year or/and only the field of the issue and get the results. Alternatively he could complete only the key words that he wants, or/and the year, or only the key words etc. The results that satisfy the user needs return in descending order of relevance with the keywords that the user has introduced. What appears on the screen is a list of the titles of the O.G.G.s that are links to the HTML pages with the highest score. Beneath the names there is a description, an indicating part of what is included in the respective page of the O.G.G. and a link for the respective pdf file of the O.G.G.. If the user clicks on any link of an html page, then the respective page of the O.G.G. appears on screen in exactly the same form as the pdf's form. Inside this page, the query words are highlighted. There are also on this page, as mentioned before, the links for navigation in previous and next page of the O.G.G. (if there is any) as well as the link to return to the initial page of the results. On the initial page of the results, the pager can be found, so that the user can go to the sequel of the results list.

4 Conclusions and future plan

The retrieval of legal documents is a recent development of Information Retrieval and Natural Language Processing for mining useful information from such documents. In this article we have presented a passage retrieval system of Greek documents in the Law domain. Emphasis was given to keep the interface as simple as possible for non-experienced users and on browsing the returned documents by the system, in such a way, that the user gets exactly the same look and feels as in the case of the printed version of the same documents. This was achieved by producing, in a fully automatic way, HTML documents with exactly the same format as their printed counterparts. The access to a retrieved document is realized through a page, the one with the highest score. Currently a user may move forwards and backwards in a document via automatically assigned links.

Several extensions of this research are currently under investigation. These include the application of natural language processing for the semi-automatic assignment of links inside a document as well as between documents. Thus the user will have the ability to jump directly to a legislation he is interested and from there, using only his mouse, could visit other related legislation or legislation that affects or is affected by the document he is visiting. In concluding, information retrieval is an active and hot research area and best practices have yet to emerge.

5 References

[Wahlgren,1999] Peter Wahlgren. The Quest for Law: Law Libraries and Legal Information Management of the Future. Jure, Stockholm, 1999.

[Simitis,1970] Spiros Simitis. Informationskrise des Rechts und Datenverarbeitung. C.F. Müller, 1970.

[trec] http://trec-legal.umiacs.umd.edu/

[clef] http://clef2012.org/

[O.G.G.] http://el.wikipedia.org/wiki/Εφημερίδα_της_Κυβερνήσεως

[Nomos] http://lawdb.intrasoftnet.com/nomos/nomos_frame.html

[Isokratis] http://www.dsanet.gr/1024x768.htm

[Digital Legal Library] http://www.nbonline.gr/

[Plati, 2012] Nomothetis: A Retrieval System in the Law Domain, BSc thesis, Athens University of Economics and Business, 2012.

[pdftohtml] http://pdftohtml.sourceforge.net/

[HtmlAsText] http://www.nirsoft.net/utils/htmlastext.html

[Lucene] http://lucene.apache.org/core/

[GreekStemmer] T.Z. Kalamboukis, (1995) "Suffix stripping with modern Greek", Program: electronic library and information systems, Vol. 29 Iss: 3, pp.313 – 321

[Java] http://en.wikipedia.org/wiki/Java_(programming_language)

[Manning, 2008] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[Rosso, 2011] Passage retrieval in legal texts, Paolo Rosso, Santiago Correa, Davide Buscaldi, Journal of Logic and Algebraic Programming, Volume 80, Issues 3–5, April–July 2011, Pages 139–153